

## Unsupervised Learning of Individuals and Categories from Images

**Stephen Waydo**

*waydo@cds.caltech.edu*

*Control and Dynamical Systems, California Institute of Technology, Pasadena,  
CA 91125, U.S.A.*

**Christof Koch**

*koch@klab.caltech.edu*

*Computation and Neural Systems, California Institute of Technology, Pasadena,  
CA 91125, U.S.A.*

Motivated by the existence of highly selective, sparsely firing cells observed in the human medial temporal lobe (MTL), we present an unsupervised method for learning and recognizing object categories from unlabeled images. In our model, a network of nonlinear neurons learns a sparse representation of its inputs through an unsupervised expectation-maximization process. We show that the application of this strategy to an invariant feature-based description of natural images leads to the development of units displaying sparse, invariant selectivity for particular individuals or image categories much like those observed in the MTL data.

### 1 Introduction ---

Neurons have been identified in the human medial temporal lobe (MTL) that display a strong selectivity for only a few stimuli (such as familiar individuals or landmark buildings) out of perhaps 100 presented to epileptic patients (Quian Quiroga, Reddy, Kreiman, Koch, & Fried, 2005; Waydo, Kraskov, Quian Quiroga, Fried, & Koch, 2006). While highly selective for a particular object or category, these cells are remarkably insensitive to different presentations (i.e., different poses and views) of their preferred stimulus. This invariant, sparse, and explicit representation of the world may be crucial to the transformation of complex visual stimuli into more abstract memories. Sparse coding as a computational constraint applied to the representation of natural images has been shown to produce receptive fields strikingly similar to those observed in mammalian primary visual cortex (Olshausen & Field, 1996). We here apply sparse coding further along the visual hierarchy: not directly to images but rather to an invariant feature-based representation of images analogous to that found

in inferotemporal cortex (Tanaka, 1997). This combination of sparseness and invariance naturally leads to explicit category representation. That is, by exposing the model to different images drawn from different categories (in this case, images of different individuals), we develop units that respond selectively to different categories.

**1.1 Related Work.** This problem is clearly distinct from the more common approach to object recognition or classification in which a labeled training set is used to learn features common to the category. These features are then extracted from unlabeled images to classify them (Barnard et al., 2003). From a pure engineering standpoint, in many settings such as recognition of objects from previously learned templates, this supervised approach is likely to be the best one. However, the MTL data suggest that the brain is capable of forming internal representations of objects in the absence of explicit supervisory signals, the issue we explore here. Further, problems such as clustering and classification of large image databases will likely benefit from at least a partially unsupervised approach, as human labeling of all images may not always be feasible. Only a few examples of truly unsupervised classification exist in the literature. The only directly comparable work is that of Sivic, Russell, Efros, Zisserman, and Freeman (2005), who address much the same computational task using very different techniques. As in our work, they first compute a feature-based (as opposed to pixel-based) representation of images, but they do so using vector-quantized scale invariant feature transform (SIFT) descriptors (Lowe, 1999) where the quantized features are obtained from a  $k$ -means algorithm applied to descriptors from sample images from their input set. By contrast, we obtain our feature-based representation using a more biologically plausible model of visual processing, the most recent extension of the HMAX model (Riesenhuber & Poggio, 1999; Serre, Oliva, & Poggio, 2007). Sivic et al. also apply a generative statistical model to the image features, using techniques developed for unsupervised topic discovery in text applied to the “words” (features) extracted from each image. An important distinction is that they found it important to restrict the number of categories (topics) searched for to the number truly present in their data sets, while our method is robust to varying numbers of input categories (and objects could in principle belong to multiple categories). Nonetheless, the essential computational approach of first building a feature-based representation of images and then learning a generative statistical model for these features is the same.

Between the extremes of fully supervised and unsupervised classification lie a number of different approaches that can be described as weakly or partially supervised, in which at least some information about the stimulus set is provided to the algorithm. Fergus, Perona, and Zisserman (2003) use an unsupervised generative learning algorithm to build representations of particular image categories, but only images from a single category are presented to the model, which is then tested in a category-versus-background

setting. Their model thus attempts to find the features common to all the images in the input set because it is known that they all come from the same category. In contrast, our model simultaneously learns representations for multiple image categories without a priori specification of the labels (or even the number of categories present). Weber, Welling, and Perona (2000) also cast the unsupervised categorization problem as emergent population coding, but again present images from only a single category at a time. The different components of their representation then correspond to different views or subcategories of the input category, and each image is explained by a single component. In principal, their method could be applied to an input set consisting of images from multiple categories, and it should distinguish between them. As with Sivic et al. (2005), however, it would be important to specify the number of categories to be identified. Dong and Bhanu (2003) present a method for image search in which the user can specify whether returned images were relevant to the search. As in this work, image features are modeled as a gaussian mixture dependent on the components (causes) present in the image, and the components of this model are estimated using unsupervised expectation maximization. Over time, a subset of images in the database is labeled through user feedback, and the system makes use of these labels to refine the category clustering.

Sparse coding as a tool for efficient representation and classification has attracted a great deal of attention in recent years, in the context of vision and elsewhere. Olshausen and Field (1996, 1997) developed the algorithm we extend here and showed that when applied to natural image patches, it generates feature selectivity much like that observed in simple cells in primary visual cortex. Hinton and Ghahramani (1997) also cast sparse representation in a generative modeling framework, but as with Olshausen and Field, they work directly at the image level. Mutch and Lowe (2006) improve the performance of the underlying vision system model we use here, in part using sparsification to enhance selectivity lower in the hierarchy. They evaluate performance in a supervised setting by training a support vector machine (SVM) for category selection. Ranzato, Poultney, Chopra, and LeCun (2006) take an energy-based approach to the unsupervised learning of sparse representations of natural images and briefly discuss its extension to a hierarchical model, though their results are at a much lower level of the visual hierarchy and so do not address categorization. Their approach, if applied to a higher level of the feature hierarchy, may produce results similar to our own. The categorization task we discuss here can be viewed as a blind source separation problem. Li, Cichocki, and Amari (2004) discuss the utility of sparse coding applied to this problem, including the aspect that the number of sources is unknown. They consider several applications, including separating speech signals and separating mixed (superimposed) drawings of faces, but not the image categorization task we discuss here.

## 2 Approach

---

We first use a biologically motivated model of hierarchical, feedforward visual processing (Riesenhuber & Poggio, 1999; Serre, Oliva, et al., 2007) to generate an image representation based on scale- and position-invariant features. This model is an extension of the Hubel and Wiesel simple-to-complex cell hierarchy. The bottom  $S_1$  layer consists of units that, like V1 simple cells, are tuned to oriented bars and edges at a variety of scales and orientations. In the next layer,  $C_1$ , each unit pools the responses of  $S_1$  units with the same preferred orientation but with small variations in position and scale, increasing the receptive field size and the invariance to these transformations and modeling complex cell behavior. Continuing up the hierarchy, each  $S_2$  unit is tuned to the activity of nearby  $C_1$  units with different feature selectivity, increasing the complexity of the unit's preferred feature, and each  $C_2$  unit pools the responses of similar  $S_2$  units over position and scale. In this way, both feature complexity and receptive field size increase in progressing up the hierarchy, until at the output layers of the model, each unit responds to the presence of a particular complex feature located anywhere in the input image. The most recent version of this model (which we use here, available at <http://cbcl.mit.edu/software-datasets/>) incorporates two parallel processing paths with somewhat different parameters for the selectivity and pooling range: one with three simple-to-complex stages terminating with layer  $C_3$  and another with two stages terminating with layer  $C_{2b}$ . We do not make use of the task-specific  $S_4$  layer that normally sits atop these layers, relying instead on the task-independent  $C_{2b}$  and  $C_3$  outputs. Despite being designed primarily to model the biological system, this model has been shown to perform on par with the state of the art in image classification tasks in a supervised setting (Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007), and even to match human performance in a rapid categorization task (Serre, Oliva, et al., 2007).

The 2000 (normalized) outputs of the  $C_{2b}$  and  $C_3$  layers of this model, which can be thought of as V4/IT neurons, were used as inputs to the sparse coding model described here. Following the approach of Olshausen and Field (1996, 1997) (but at the feature rather than the pixel level), we assume that, for a given image, these inputs  $u \in \mathbb{R}^n$  are a noisy linear function of some set of causes  $v \in \mathbb{R}^m$ , or

$$u = Gv + \xi, \tag{2.1}$$

where the columns of the matrix  $G$  form a basis for the outputs  $u$  and  $\xi \in \mathbb{R}^n$  is zero-mean gaussian noise with covariance  $\lambda I$ . In our case  $m \ll n$ . The causes  $v$  change from one image to the next, and the goal of learning is to find a set of weights  $G$  such that equation 2.1 provides an accurate description of the input data for sparse and independent causes  $v$ .

To enforce the constraint that the causes be sparse and independent, we impose a sparse prior distribution, that is, we set

$$f(v) \propto \prod_{i=1}^m \exp(S(v_i)), \quad (2.2)$$

where  $v_i$  is the  $i$ th element of  $v$  and  $S(v_i)$  is defined such that the resulting distribution is sparse. The exponential form of the prior is chosen simply for mathematical convenience. For simplicity, we omit the proportionality constant required to make this distribution integrate to 1 (this constant would drop out of the forthcoming optimization, so there is no loss of generality). In Olshausen and Field (1997), where this strategy was used to develop a visual cortex-like sparse code for natural images, the sparse prior  $S$  followed a Cauchy distribution. Because we seek to develop units that respond in a more-or-less binary fashion (i.e., most responses are close to 0, while a few will be close to 1), we instead use a weighted sum of two gaussians with means 0 and 1,

$$\exp(S(v_i)) = \frac{1-t}{\sqrt{2\pi}\sigma} e^{-\frac{v_i^2}{2\sigma^2}} + \frac{t}{\sqrt{2\pi}\sigma} e^{-\frac{(v_i-1)^2}{2\sigma^2}}, \quad (2.3)$$

where  $t$  is the (small) desired probability of a strong response. We further constrain the rates  $v_i$  to be positive.

In their work on applying sparse coding principles to representation in V1, Olshausen and Field (1997) maximized the average log likelihood that the model would generate the observed input data, that is,

$$\begin{aligned} \mathcal{F}(G, v(u)) &= \langle \ln f(v(u), u) \rangle \\ &= \left\langle -\frac{1}{2\lambda} \|u - Gv(u)\|^2 + \sum_{i=1}^m S(v_i(u)) \right\rangle. \end{aligned} \quad (2.4)$$

The first term of this cost function penalizes a mismatch between the true input  $u$  and the modeled input  $Gv$ , and so expresses how well the current model represents the input set. The second term enforces the sparseness constraint embodied in  $S$ . As discussed in Olshausen and Field (1997), however, this optimization has a trivial solution in which the elements of  $G$  grow without bound. Roughly speaking, one can always improve the cost by uniformly decreasing  $v$  and increasing  $G$ . This problem was alleviated by periodically normalizing the columns of  $G$  to maintain a desired output variance. We instead express the constraint that large weights are unlikely by including a zero-mean gaussian prior distribution on the elements  $g_{ij}$  of

G. This leads to a quadratic penalty on these weights in  $\mathcal{F}$ ,

$$\begin{aligned}\mathcal{F}(G, v(u)) &= \langle \ln f(v(u), u, G) \rangle \\ &= \left\langle -\frac{1}{2\lambda} \|u - Gv(u)\|^2 + \sum_{i=1}^m S(v_i(u)) \right\rangle - \frac{1}{2\gamma} \sum_{i=1}^n \sum_{j=1}^m g_{ij}^2,\end{aligned}\tag{2.5}$$

where  $\gamma$  is the variance of the prior distribution.

We maximize  $\mathcal{F}$  using an iterative expectation-maximization (EM) algorithm. In the E step, for each input  $u$ , we compute the most likely cause  $v(u)$ . Performing gradient ascent with respect to  $v$  (for a particular  $u$ ) we obtain the differential equation,

$$\dot{v} = \frac{1}{\lambda} G^T (u - Gv) + S'(v),\tag{2.6}$$

where  $S'$  is the derivative of  $S$  with respect to its argument. The simulation is carried out until it reaches equilibrium, at which point  $v$  is a (local) maximizer of  $\mathcal{F}$ . This system has a recurrent neural network implementation in which the nonlinear dynamics of the output neurons are defined by  $S'$ . The recurrent feedback ( $-G^T G$ ) term introduces competition between output units that represent similar inputs, producing winner-take-all behavior. This stage of the optimization can be viewed as computing the set of basis functions that best represent the input, subject to the sparseness constraint imposed by  $S$ .

In the M step, we compute the optimal  $G$  for the current  $v(u)$ . Taking the derivative of  $\mathcal{F}$  with respect to  $G$ , setting this expression equal to zero, and solving for  $G$ , we obtain the update rule:

$$G \rightarrow \langle uv^T \rangle \left( \frac{\lambda}{\gamma} I + \langle vv^T \rangle \right)^{-1}.\tag{2.7}$$

Because  $\frac{\lambda}{\gamma} I$  is positive definite and  $\langle vv^T \rangle$  is positive semidefinite, their sum is positive definite and thus nonsingular, so this learning rule is always well defined and yields the globally optimal  $G$  for the current  $v(u)$ . This rule is a significant extension of the method, as the large M step results in much faster convergence of the EM algorithm than the incremental rule presented in Olshausen and Field (1997).

If online, incremental learning is desired (as would arise in a biological context), we can instead implement a gradient-ascent update rule for  $G$ .

For a particular input  $u$  and cause  $v(u)$  computed as above, gradient-ascent learning results in the update rule

$$G \rightarrow G + \eta \left( \frac{1}{\lambda} (u - Gv(u))v(u)^T - \frac{1}{\gamma} G \right), \quad (2.8)$$

where  $\eta$  is a small, positive learning constant. This is a Hebbian learning rule (with decay) between the cause  $v(u)$  and the reconstruction error  $(u - Gv(u))$  and can be realized locally in the network implementing equation 2.6. This is essentially the rule proposed by Olshausen and Field (1997), though we extend that work to include the decay term used here. While it results in much slower convergence than the batch rule (see equation 2.7), it has the advantage of a local, biologically plausible implementation.

### 3 Results

---

This algorithm was tested on a data set consisting of gray-scale frontal facial images of different individuals obtained from the Caltech-256 data set (Griffin, Holub, & Perona, 2006). Though the backgrounds vary slightly from image to image, these images are fairly well structured and could be viewed as the output of an attentional selection and segmentation process. Training was performed using 10 different images of each individual, with 10 different images of the same individuals reserved for testing. We performed experiments with 4 to 10 different individuals in the input set. The number of inputs to the network (outputs of the preprocessing system) was  $n = 2000$ , and the number of output units was  $m = 15$ . The sparseness and variance parameters of the sparse prior distribution (see equation 2.3) were  $t = 0.05$  and  $\sigma^2 = 0.04$ , and the noise covariance parameter, which expresses the relative weight between the sparse prior and the reconstruction error, was  $\lambda = 10$ . The parameter imposing the penalty on synaptic weights was  $\gamma = 100$ . Learning was carried out using the batch learning rule (see equation 2.7) to minimize computation time, but the results are no different from those obtained using the incremental learning rule (see equation 2.8).

Through training, units developed selective, invariant responses to particular individuals, which we quantify in two ways. First, we consider each unit individually as a single category classifier in a receiver operating characteristic (ROC) framework. For each category in each session, we find the trained unit displaying the strongest selectivity and plot the detection rate versus the false alarm rate for the unit taken as a linear binary classifier for a range of thresholds. A unit responding randomly to different individuals will have an ROC curve close to the diagonal, while a unit responding selectively to a single individual will have a curve far from the diagonal, with an area under the curve close to 1. We then compute the ROC equal-error accuracy, that is, the accuracy at which  $p(\text{true positive}) = 1 - p(\text{false$

positive) and use this as our selectivity metric. This accuracy corresponds to the intersection of the ROC curve with the diagonal from (0, 1) to (1, 0). The ROC measure best quantifies the degree to which a sparse, invariant representation of each category has emerged. Second, we use all selective units together to perform the multicategory classification task in a weakly supervised way. In this setting, each trained unit is labeled with the category for which it is most selective (according to the ROC metric), and then each test image is assigned to the category of the unit with the strongest response. We then form the confusion matrix for this system; the average of the diagonal of this matrix is the overall classification accuracy. This measure indicates the level of confusion between categories represented by different units. As a baseline for comparison, we also evaluated the performance of the first 15 principal components of the 2000 inputs against these metrics. Finally, we found the best performance we could achieve using a supervised SVM classifier, which provides a reasonable upper bound on achievable performance and an objective measure of the task difficulty.

Figure 1 depicts the responses of two selective units (the best and a more typical unit) from a single training session with 10 different individuals in the input set. On the left we show 20 of the images that produced the strongest responses in the unit, with every second image omitted to better span the image set. On the right is a histogram of all responses of the unit, with responses to the preferred individual in black and responses to all others in white. Inset in the histogram is the ROC curve for the unit (solid line) and the best ROC curve from principal component analysis (PCA; dashed line). The mean ROC accuracy (that is, the average ROC accuracy of the best unit for each category) for this run was 91%, and the ROC accuracies for the two units shown were 100% (see Figures 1a and 1b) and 90% (see Figures 1c and 1d). The semisupervised 10-way classification accuracy of the sparse network was 56%. PCA yielded a mean ROC accuracy of 78% and a semisupervised classification accuracy of 37%. Additionally, in contrast to the responses of the sparse units depicted in Figure 1, the responses of the principal components were unimodal and so did not clearly indicate the presence of a category in the same way as the sparse units (which is reflected in the poor semisupervised accuracy).

We repeated this experiment 50 times for each number of different individuals, each time starting with different random initial conditions (initial synaptic weights), using a different random subset of the 17 individuals for whom we had at least 20 pictures, and using different random subsets for training and testing. Figure 2 summarizes the results and compares them to those obtained from the top 15 principal components of the inputs and the limit suggested by supervised classification. Performance according to the ROC metric did not vary significantly with the number of people presented, indicating that in all cases, units emerged that responded selectively to each individual. The mean ROC accuracy across all 350 trials was 91.3%, compared to 96.6% for a binary SVM and 80.4% for PCA. Performance



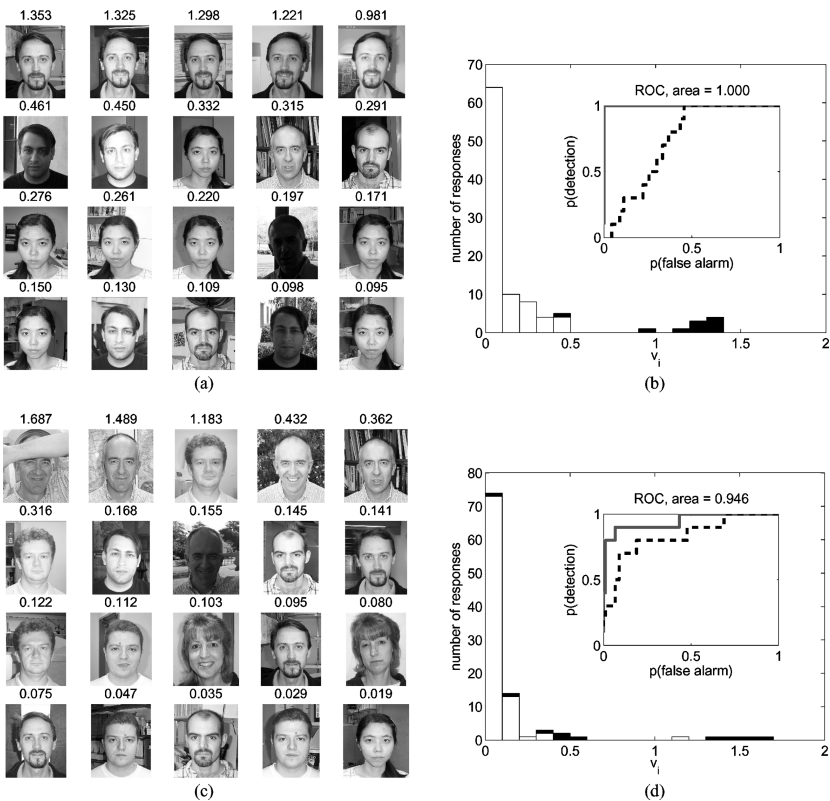


Figure 1: Responses of two selective units (out of 15) after the unsupervised category learning. (a, c) Images that evoked the top responses, with the activation level above each image. Every second image is omitted for clarity. (b, d) Response histograms. The  $x$ -axis is the activation level; the  $y$ -axis is the number of responses (100 total) evoking a response at that level. The responses to the preferred person are in black; responses to all other images in white. Insets: ROC curves. The solid line is the ROC curve for the selected unit (exactly along the vertical and horizontal axes in  $b$ ); the dashed line is the ROC curve for the best principal component.

according to the semisupervised metric did decline as the number of people in the input set increased, dropping from a mean of 85.5% to 64.2% as the number of people increased from 4 to 10. This is in all cases significantly better than the PCA performance, which decreased from 58.1% to 41.1%. This decline is not unexpected, because as more categories are presented, it becomes more likely that in addition to the “correct” unit responding to a given image, some other unit will spuriously respond strongly.

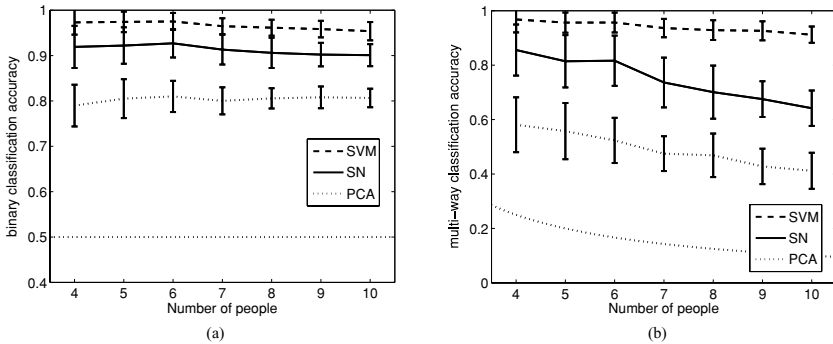


Figure 2: Classification accuracy (mean  $\pm$  SD) as a function of the number of people in the input set. Solid line: sparse coding network; dashed line: SVM (supervised) classifier; dotted line: PCA. The dotted line without error bars depicts chance performance. (a) ROC equal-error accuracy for binary classification. (b) Semisupervised multiway classification accuracy.

#### 4 Other Data Sets

We have applied this algorithm to several other data sets to explore its robustness to greater variation in object presentation and its performance in categorizing more broadly defined categories. To evaluate robustness to more widely varied facial images, we collected images of several celebrities on the Web (Jennifer Aniston, Halle Berry, George Clooney, and Matt Damon) as in the original electrophysiological study by Quian Quiroga et al. (2005). Images were selected that contained reasonably frontal views of faces and cropped to contain only the face, so the overall composition was similar to the images used in Figure 1 (the images cannot be shown here due to insurmountable copyright issues). These images contained substantially more variation in pose, facial expression, hairstyle, and background than the images used in Figure 1 and were much more difficult to classify even using supervised methods. We again performed 50 trials using 10 images for training and 10 for testing, randomizing over initial weights and which images were used for training and testing. The average ROC accuracy was 77.4%, compared to an average supervised SVM accuracy of 84.3%. Relative to the benchmark of supervised classification, then, performance was essentially the same as before.

To explore performance in a more general categorization task, we applied the algorithm to images from four categories drawn from the Caltech-256 data set: airplanes, cars, motorbikes, and faces. More images were available for each category than in the face discrimination task; the training and testing sets consisted of 40 images each. The number of output units was  $m = 10$ ; all other network parameters were the same as before. Again, the

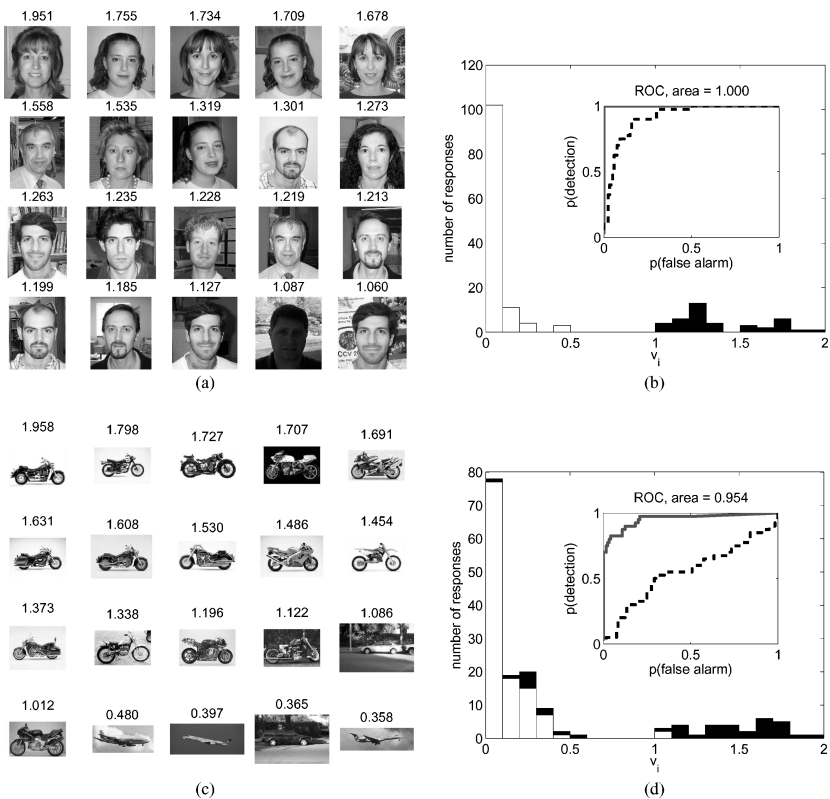


Figure 3: Responses of two selective units (out of 10) after the unsupervised category learning. (a, c) Images that evoked the top responses, with the activation level above each image. Every second image is omitted for clarity. (b, d) Response histograms. The  $x$ -axis is the activation level; the  $y$ -axis is the number of test images (160 total) evoking a response at that level. Responses to the preferred category are in black; responses to all other images are in white. Insets: ROC curves. The solid line is the ROC curve for the selected unit (exactly along the vertical and horizontal axes in  $b$ ). The dashed line is the ROC curve for the best principal component.

model successfully learned sparse, invariant representations for the four input categories, with an average ROC accuracy over 10 trials of 89.8% compared to an SVM accuracy of 97.4%. Figure 3 depicts the responses of two selective units (the best and a more typical unit) after training. ROC accuracies for the two units shown were 100% (see Figures 3a and 3b) and 88% (see Figures 3c and 3d). Interestingly, in this setting in which faces were part of the input set but no particular face was presented often, the

model developed a generic representation for “face” rather than differentiating among faces as seen above when only faces comprise the training set.

## 5 Conclusions

---

These results show that a sparse, invariant, and explicit representation of individuals can emerge from an unsupervised learning algorithm with only the simple constraint that it should provide an accurate reconstruction of its inputs using sparse causes. Crucial to the success of this algorithm is that it is applied to an image representation that provides invariance to unimportant transformations (e.g., scale and position). It is our belief that the general architecture of representing sensory inputs with invariant features and then learning sparse representations of the inputs in this feature space naturally leads to category learning independent of the exact models used for each of these stages (provided enough information is preserved in the feature representation stage). A future challenge is to examine how robust this framework is to different input representations and sparse coding models. We are encouraged in this regard by the work of Sivic et al. (2005), who apply very different algorithms in a similar conceptual framework with some success.

The very weak variation of ROC classification accuracy with number of input categories suggests that the method robustly generates sparse, invariant representations of its inputs even as the size of the input set scales up. However, it is likely that accurately classifying larger numbers of categories will require a more sophisticated underlying feature model that more tightly groups images from the same category in feature space, as even the supervised SVM accuracy begins to drop off with faces from 10 or more individuals in the input set.

We have shown here that the same model successfully employed by Olshausen and Field to model V1 can also be fruitfully applied to a much higher level of vision, and so it is reasonable to expect that it could be applied throughout the visual hierarchy to provide the needed performance improvement. The primary obstacle to this approach is one of available computational resources. The intermediate layers of the vision model used here consist of millions of simulated neurons, and so the model is tractable only because these neurons operate in a purely feedforward fashion. By contrast, interactions among neurons in the same layer are crucially important to our sparse coding scheme, and so a more efficient means for computing the equilibrium of equation 2.6 (and thus computing the representation) would be required. Another interesting avenue for further research is the implication of this theoretical work on the time course of responses and density of highly selective neurons in the human brain.

## Acknowledgments

---

This work was supported by grants from NIMH, NSF, ONR, DARPA, the Mathers Foundation, and a Fannie and John Hertz Foundation fellowship to S.W. Thomas Serre and Minjoon Kouh of MIT provided invaluable assistance in the setup and operation of the underlying vision model. We thank Richard Murray, Pietro Perona, Jerry Marsden, Tomoso Poggio, Bruno Olshausen, and the members of klab for valuable comments on this work.

## References

---

- Barnard, K., Duygulu, P., Freitas, N. de, Forsyth, D., Blei, D., & Jordan, M. (2003). Matching words and pictures. *Journal of Machine Learning Research*, 3, 1107–1135.
- Dong, A., & Bhanu, B. (2003). A new semi-supervised EM algorithm for image retrieval. In *Proceedings of the IEEE Computer Science Conference on Computer Vision and Pattern Recognition 2003*. Washington, DC: IEEE.
- Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Computer Science Conference on Computer Vision and Pattern Recognition 2003*. Washington, DC: IEEE.
- Griffin, G., Holub, A., & Perona, P. (2006). *The Caltech 256* (Tech. Rep. CNS-TR-2007-001). Pasadena: California Institute of Technology.
- Hinton, G., & Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Phil. Trans. R. Soc. Lond. B*, 352, 1177–1190.
- Li, Y., Cichocki, A., & Amari, S. (2004). Analysis of sparse representation and blind source separation. *Neural Computation*, 16, 1193–1234.
- Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*. Washington, DC: IEEE.
- Mutch, J., & Lowe, D. (2006). Multiclass object recognition with sparse, localized features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2006*. Washington, DC: IEEE.
- Olshausen, B., & Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381, 607–609.
- Olshausen, B., & Field, D. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23), 3311–3325.
- Quiñero-Ruano, R., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435, 1102–1107.
- Ranzato, M., Poultney, C., Chopra, S., & LeCun, Y. (2006). Efficient learning of sparse representations with an energy-based model. In H. I. Jordan, Y. LeCun, & S. A. Solla (Eds.), *Advances in neural information processing*, 19. Cambridge, MA: MIT Press.
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11), 1019–1025.
- Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Science*, 104(15), 6424–6429.

- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), 411–426.
- Sivic, J., Russell, B., Efros, A., Zisserman, A., & Freeman, W. (2005). *Discovering object categories in image collections* (Tech. Rep. No. MIT-CSAIL-TR-2005-012). Cambridge, MA: MIT, Computer Science and Artificial Intelligence Laboratory.
- Tanaka, K. (1997). Mechanisms of visual object recognition: Monkey and human studies. *Current Opinion in Neurobiology*, 7, 523–529.
- Waydo, S., Kraskov, A., Quiroz Quiroga, R., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience*, 26(40), 10232–10234.
- Weber, M., Welling, M., & Perona, P. (2000). Towards automatic discovery of object categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2000*. Washington, DC: IEEE.

---

Received March 20, 2007; accepted July 16, 2007.